AD_____

Award Number: W81XWH-11-1-0136

TITLE: Automated Discovery of Long Intergenic RNAs Associated with Breast Cancer Progression

PRINCIPAL INVESTIGATOR: Matthew Iyer

CONTRACTING ORGANIZATION: The University of Michigan
Ann Arbor, MI 48109

REPORT DATE: February 2012

TYPE OF REPORT: Revised Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
February 2012	Revised Annual Summary	15 January 2011 - 14 January 2012

4. TITLE AND SUBTITLE

Automated Discovery of Long Intergenic RNAs Associated with Breast Cancer Progression

5a. CONTRACT NUMBER

5b. GRANT NUMBER

W81XWH-11-1-0136

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Matthew Iyer

E-Mail: diannal@umich.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The University of Michigan
Ann Arbor, MI 48109

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Of the ~40,000 women in the United States that die from breast cancer every year, almost half of them might have once thought they were cured. While treatment of early stage breast cancer can be curative, up to 30% of node-negative and 70% of node-positive breast cancers will relapse. Therefore, risk stratification and surveillance of breast cancer is of paramount importance. Prognostic markers such as estrogen and progesterone receptor, ERBB2, and comedo necrosis currently guide clinical decisions, but these markers often fail to estimate the true risk of relapse for many patients. This results in frequent overtreatment of indolent cancer and undertreatment of high-risk disease. Over the past decade gene expression microarrays have facilitated the development of prognostic tests, but Next Generation Sequencing of cancer transcriptomes (RNA-seq) technologies can provide more information at higher accuracy. However, current RNA-seq analysis tools cannot resolve a significant portion of the data emanating from cancer transcriptomes, and it is precisely this data that could lead to the discovery of novel genetic aberrations and/or therapeutic targets. Therefore, we hypothesize that transcriptome sequencing can elucidate novel transcriptional aberrancies in breast cancer that may affect or predict disease prognosis. We will test this hypothesis with the following specific aims: 1) we will employ a combined alignment and assembly approach to increase the sensitivity of current analysis methods, 2) use this approach to detect novel transcriptional aberrancies in breast cancer, and 3) predict functional relationships between novel transcripts and known prognostic markers. Methods to detect and characterize these transcripts could lead to discovery of oncogenic mutations, delineate new molecular subtypes of breast cancer, and/or identify novel therapeutic targets for breast cancer treatment.

15. SUBJECT TERMS Breast cancer; bioinformatics; long non-coding RNA; lncRNA; lincRNA; gene fusions; transcriptome sequencing; high-throughput sequencing; RNA-seq

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	19	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Matthew Iyer
M.D./Ph.D. Candidate, University of Michigan
Project report
February 14, 2012

Introduction

Our lab has sequenced transcriptomes of breast cancer tissues and cell lines using the Illumina Genome Analyzer II and Illumina Hi-Seq 2000 platforms. **We hypothesize that transcriptome sequencing can elucidate novel transcriptional aberrancies in breast cancer that may affect or predict disease prognosis.** To test this hypothesis, the P.I. proposed a work plan consisting of three Specific Aims: (1) to develop a bioinformatics approach that comprehensively annotates breast cancer transcriptomes, (2) to detect transcripts that associate with cancer stage and subtype, and (3) to develop methods to correlate novel transcripts with known prognostic genes. Progress on each of these aims and updates to the work plan are detailed in this annual report.

Body

Over the past year the P.I. has made substantial progress towards completion of the three Specific Aims proposed in the original work plan. Major accomplishments include (1) development of an automated RNA-Seq data analysis pipeline for analyzing breast cancer RNA-Seq data (*manuscript in preparation*), (2) development and publication of an algorithm for detecting gene fusions in RNA-Seq data [1], and (3) discovery of outlier long non-coding RNAs (lncRNAs) and fusion transcripts in breast tumor RNA-Seq data [2]. Detailed progress is reported below for each of the proposed Specific Aims:

Specific Aim 1

Specific Aim 1 is to develop a bioinformatics approach that uses RNA-Seq data to comprehensively annotate breast cancer transcriptomes. We proposed to (A) develop an infrastructure that uses the TopHat software tool while simultaneously (B) applying *de novo* assembly strategies to the unaligned reads. Utilizing *de novo* strategies allows detection of complex genomic aberrations such as indels, viral sequences, and gene fusions [3,4].

During the 2011 award period, the P.I. developed a tool to automatically determine the cDNA fragment length from sequencing data by aligning sequences to the reference transcriptome and profiling the fragment length distribution of the high-quality alignments to determine the mean and standard deviation of the insert size. The insert size parameters were used to align reads using TopHat [5] to discover novel splice junctions arising from isoforms and intergenic transcripts. For each sample we generate a BigWig [6] format file for viewing the coverage in genome browser tools such as UCSC or IGV. We visualized novel splice variants in the genome browser and confirmed that TopHat can discover such events. The entire software infrastructure was packed into an automated pipeline called Oncoseq that can analyze large cohorts of RNA-Seq data in parallel on a supercomputing cluster. We ran the infrastructure on our cohort of breast tissue samples as well as 51 breast cell lines. The analysis results were then subjected to *de novo* assembly algorithms to discover novel transcripts representing either unannotated genes or novel somatic mutations such as gene fusions. To this end the P.I. developed and published a novel algorithm called ChimeraScan to facilitate the discovery and validation of gene fusion events in RNA-Seq data [1]. This algorithm subsequently aided in the exciting discovery of recurrent gene fusions in breast cancer involving the *NOTCH* and *MAST* families of kinases [2]. For the upcoming 2012 award period the P.I. will be further analyzing the RNA-Seq data to comprehensively annotate the breast cancer transcriptome and define lncRNAs aberrantly expressed in the disease. The P.I. has completed preliminary transcriptome assemblies for breast cancer that have been used to nominate novel transcripts that are outliers in cancer (Figure 1). Some of these novel transcripts have been validated by qRT-PCR using intron-spanning primers, confirming that the splice junction predictions are accurate (Figure 2). The transcriptome assembly algorithm is still under development and will be further refined and improved in 2012.

Long non-coding RNAs dramatically upregulated in breast tumors

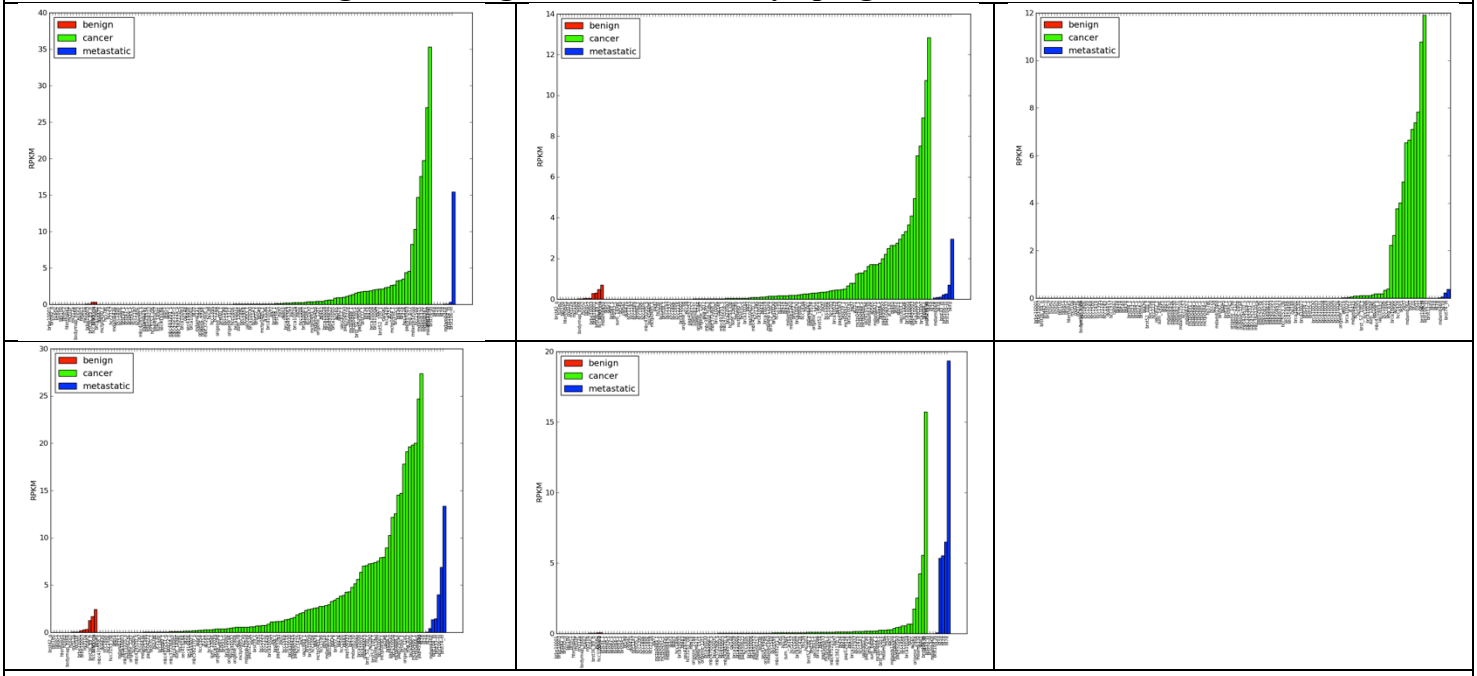

Figure 1: RNA-Seq expression profiles of five intergenic long non-coding RNA loci nominated as breast cancer outlier genes. These lncRNAs are dramatically upregulated in breast tumors and metastases.

qRT-PCR validation of lncRNA predictions

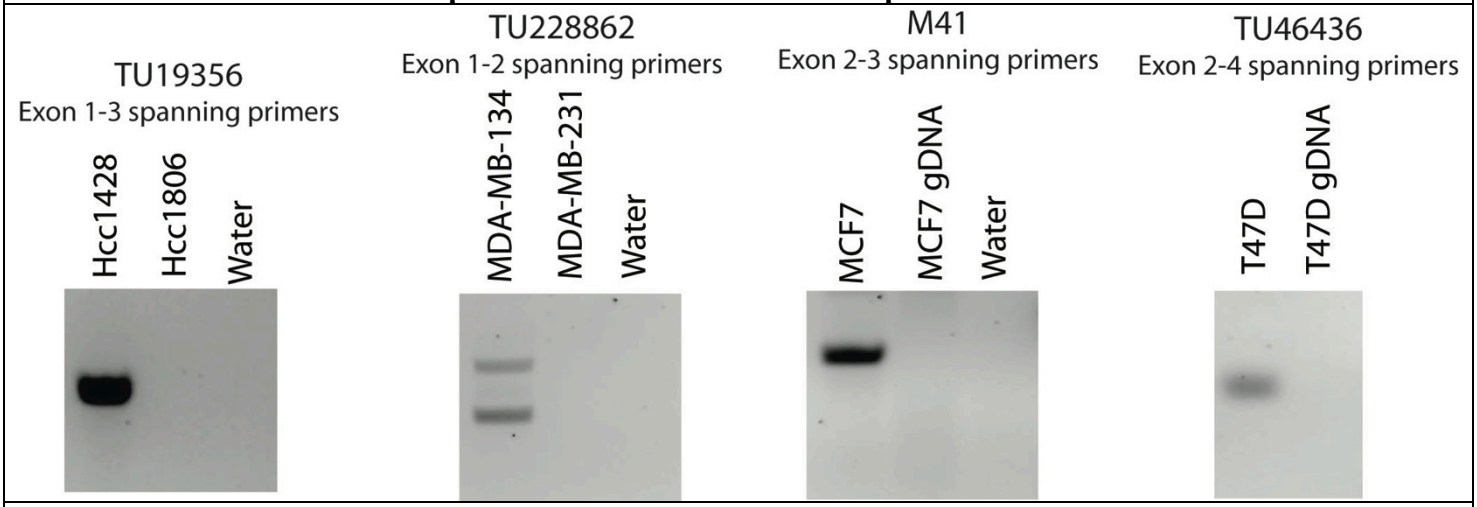

Figure 2: qRT-PCR validation of four different long non-coding RNAs nominated by de novo reconstruction of the transcriptome in a preliminary cohort of breast tissue samples.

Specific Aim 2

Specific Aim 2 proposed to apply differential expression analyses to detect aberrant transcripts associated with cancer stage and subtype. During the 2011 award period the P.I. developing software to construct gene expression matrices from many samples and verified that ER/PR and ERBB2 were outliers in mutually exclusive sample subsets. Additionally, the P.I. evaluated several publicly available software packages including *DESeq (http://www-huber.embl.de/users/anders/DESeq), DEG-Seq (http://www.bioconductor.org/packages/2.6/bioc/html/DEGseq.html), edgeR (http://bioconductor.org/packages/2.10/bioc/html/edgeR.html), and cuffdiff* (*http://cufflinks.cbcb.umd.edu/index.html*) [7,8]. We determined that *DESeq* was most robust for determining

gene-level differential expression events and will be using this tool in subsequent analyses. Transcripts that are highly expressed in a minority of cancer tissues may not be detectable by standard differential expression analysis but could represent novel cancer subtypes. These transcripts are often known as "outliers" because they only appear in a subset of the samples in a given cohort. Outlier detection algorithms have been used successfully to detect aberrant transcripts in prostate cancer, including the *TMPRSS2-ERG* gene fusion; however, these algorithms were designed for microarray gene expression data and not RNA-Seq [9]. Accordingly, the P.I. has designed, implemented, and tested a preliminary version of an RNA-Seq outlier analysis algorithm on breast cancer tissue data. The P.I. plans to publish this algorithm as an R package and submit a manuscript to a bioinformatics journal during the award period. Intriguingly, the algorithm nominated a set of five long RNA transcripts in cancer that did not appear in any normal tissues or cell lines (Figure 1). One of these transcripts, named *M41*, was expressed across multiple tissues (Figure 3), suggesting that it may play a role in multiple cancer types.

In summary, the ***P.I. applied the novel transcript discovery pipeline to preliminary breast tumor RNA-Seq data and nominated five breast cancer outliers that are long non-coding RNAs*** (Figure 1). Four of the five outliers were validated by qRT-PCR (Figure 2). One of these long non-coding RNAs, termed *M41* based on existing annotations, was highly expressed in breast, lung, and prostate cancer (Figure 3). In 2012, the P.I. will continue working on the plan for Specific Aim 2, with the goal of submitting a manuscript on the outlier analysis algorithm. Many of the items in the work plan are still in progression, and the P.I. anticipates

The M41 lncRNA is highly expressed in breast, lung, and prostate cancer

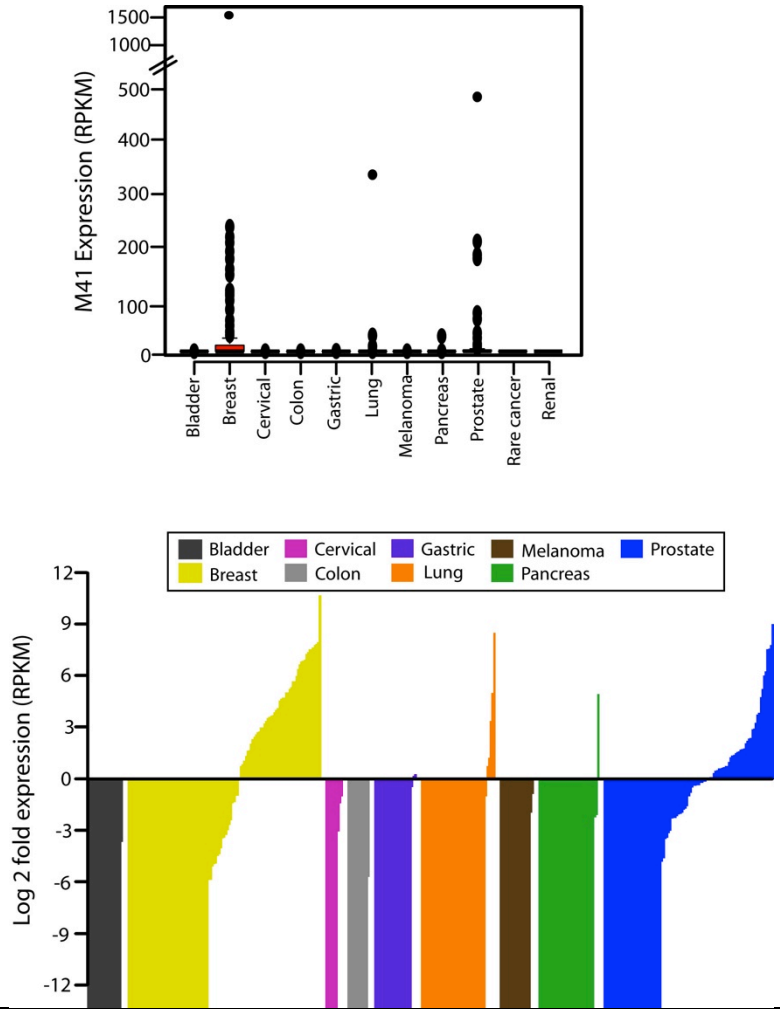

Figure 3: (upper panel) Boxplot of M41 lncRNA expression across multiple tissue types reveals that M41 is highly expressed in subsets of breast, lung, and prostate cancer samples. (lower panel) Normalized RPKM values show M41 levels across all samples.

completing the majority of these tasks in 2012.

Specific Aim 3

Specific Aim 3 proposes to develop a method that detects associations between novel transcripts and known prognostic genes. During the 2011 award period the P.I. began developing and/or implementing existing software packages for clustering and correlating novel transcripts with known protein coding genes. This work is still preliminary and under development. The P.I. expects to make significant progress on this aim during the 2012 award period.

Summary of progress towards completion of work plan

Timeline: Automated discovery of novel transcripts associated with breast cancer progression		Months 1-6	Months 7-12	Months 13-18	Months 19-24	Months 25-30	Months 31-36
Specific Aim 1: develop a bioinformatics approach to comprehensively annotate breast cancer transcriptomes							
A1	Profile insert size distributions across samples	X					
A2	Align reads using TopHat in a single sample	X					
A3	Visualize results on the UCSC Genome Browser		X				
A4	Confirm presence of novel events by manual inspection		X				
A5	Parallelize the infrastructure		X				
A6	Run on entire cohort						
A7	Pool together high quality biological replicates						
B1	Obtain/install assembly tools	X					
B2	Simulate short-reads to validate assembly approach		X				
B3	Apply assembly to unaligned reads from TopHat		X				
B4	Integrate results into a comprehensive transcriptome model						
B5	Experimentally validate novel transcripts						
Specific Aim 2: Characterize aberrant transcripts associated with clinical progression and subtype							
1	Convert raw alignment output to expression matrix format for DEG-seq	X					
2	Obtain, install, evaluate DEG-seq		X				
3	Test for differential expression between cancer stages (both cell lines and tissues)						
4	Test for differential expression between clinical subtypes						
5	Validate significant novel results using qRT-PCR						
6	Obtain, install, and evaluate Cufflinks on cell lines	X	X				
7	Validate isoform switching events in cell lines						
8	Determine set of transcripts with isoform switching associated with cancer stage and clinical subtype; validate tissue results						
9	Implement the Genomic Outlier Profile Analysis (GOPA) method for cancer outlier detection	X					
10	Apply GOPA to further characterizes breast cancer subtypes						
Specific Aim 3: To associate novel transcripts with prognostic gene expression 'signatures'							
8	Develop a clustering method to group co-expressed nearby loci						
9	Apply spatial clustering method to propose co-regulated transcript groups						
10	Explore clustering paradigms for the purpose of associating novel transcripts with prognostically-relevant genes	X					
11	Compare novel transcripts to genes in the 21-gene RS signature						
12	Assess whether novel transcripts are associated with histologic grade						

Key research accomplishments

- The P.I. developed and published a new software tool, ChimeraScan, which enables de novo detection of gene fusions in RNA-Seq data [1]. The manuscript describes the fusion detection methodology in detail and has been included the Appendices of this report. The ChimeraScan software is hosted online (http://chimerascan.googlecode.com) where it has been downloaded over 270 times by potential users. The P.I. continues to maintain and improve ChimeraScan and anticipates making additional software releases during the 2012-2013 award period.
- The P.I. has developed an infrastructure called Oncoseq for processing vast amounts of RNA-Seq data. The Oncoseq software package is now being used to analyze the breast cancer transcriptomes for discovery of novel transcripts. The package itself is versatile and capable of analyzing RNA-Seq data for a variety of other projects as well. We are currently preparing a manuscript for publication in a bioinformatics journal.
- The Oncoseq infrastructure has been used to process 207 RNA-Seq libraries from breast cancer cell lines and tissues. The results of this analysis have been used in the discovery of novel classes of gene fusions in breast cancer, including *NOTCH* and *MAST* kinase gene fusions [2]. The manuscript discussing these findings has been included in the Appendices of this report.

Reportable outcomes

Toward the fulfillment of Specific Aim 1B, ***the P.I. developed a software package for de novo assembly of gene fusion breakpoints from paired-end RNA-Seq data***. The tool, ChimeraScan [1], is now publicly available at http://chimerascan.googlecode.com and published as an Application Note at the journal Bioinformatics (see article in Appendices). In addition, the ***P.I. applied ChimeraScan towards the discovery of gene fusions and was co-author on a recent publication that nominated the MAST and NOTCH kinases as recurrent families of genetic aberrations in triple negative breast cancer [2]*** (see article in Appendices).

During the grant period through March 2012, the P.I. has completed significant coursework, examinations, and research training (Table 1). The P.I. has completed the all of the coursework required by the Ph.D. program in Bioinformatics at the University of Michigan, and has thus far held two dissertation committee meetings. Committee meetings have been planned for April 2012, October 2012, and March 2013.

Event	Update
Clinical Competency Exam	Completed
Completion of basic medical coursework	Completed
USMLE Step1 Medical Licensing Exam	Completed
Bioinformatics Preliminary Exam	Completed
Dissertation Committee Assembled	Completed (July 2010)
Dissertation Committee Meeting #1	Completed (Nov 2010)
Completion of BGP coursework	Completed (Dec, 2010)
Dissertation Committee Meeting #2	Completed (June, 2011)
Dissertation Committee Meeting #3	Scheduled (April, 2012)
Dissertation Committee Meeting #4	Scheduled (October, 2012)
Dissertation Committee Meeting #5	Scheduled (March, 2013)
MSTP Annual Conference	Annually
Lab Meetings, Journal Club	Weekly until graduation
Attendance at national meetings	Annually
MSTP clinical clerkship	Weekly beginning January 2013
Department and program seminars	As appropriate

Table 1: Training Plan Timeline

Date	Event
2011 (Oct 15-18)	Poster Presentation**, American Association for Cancer Research (AACR) Translation of the Cancer Genome**, San Francisco, CA
2011 (Oct 10-12)	Poster Presentation**, Cell Symposium on Regulatory RNAs**, Chicago, Illinois
2011 (April 2-5)	Mini-Symposium**, AACR Annual Meeting**, Orlando, Florida.
2011 (March 22)	Poster Presentation**, Prostate Cancer SPORE Meeting**, Fort Lauderdale, Florida

Table 2: Conferences and/or symposia

The P.I. presented at four conferences and/or symposia from January 2011 through February 2012 (Table 2). Additionally, the P.I. presented at weekly lab meetings, journal clubs, the 2011 Medical Scientist Training Program (MSTP) retreat, and a Bioinformatics student seminar called BISTRO.

1. **Iyer MK**, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011 Oct 15;27(20):2903-4.

2. Robinson DR, Kalyana-Sundaram S, Wu YM, Shankar S, Cao X, Ateeq B, Asangani IA, **Iyer M**, Maher CA, Grasso CS, Lonigro RJ, Quist M, Siddiqui J, Mehra R, Jing X, Giordano TJ, Sabel MS, Kleer CG, Palanisamy N, Natrajan R, Lambros MB, Reis-Filho JS, Kumar-Sinha C, Chinnaiyan AM. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. Nat Med. 2011 Nov 20;17(12):1646-51.

Conclusions

In 2011, the P.I. made significant progress towards completion of the Specific Aims proposed. He was an author on two manuscripts that focused on breast cancer samples, presented at national scientific meetings, made significant progress towards finishing his Ph.D., and made several discoveries of intergenic lncRNAs differentially expressed in breast tumors. In the coming year, the P.I. will gain enormous insight into the transcriptomic complexity of breast cancer as he moves from primary data analysis towards interpretation, and leverages the unique RNA-Seq data set in order to examine patterns of gene expression across subtypes of breast cancer. The P.I. expects to make important contributions to the field by identifying novel gene fusions, mutations, and lncRNAs implicated in breast cancer progression.

References

1. **Iyer MK**, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011 Oct 15;27(20):2903-4.

2. Robinson DR, Kalyana-Sundaram S, Wu YM, Shankar S, Cao X, Ateeq B, Asangani IA, **Iyer M**, Maher CA, Grasso CS, Lonigro RJ, Quist M, Siddiqui J, Mehra R, Jing X, Giordano TJ, Sabel MS, Kleer CG, Palanisamy N, Natrajan R, Lambros MB, Reis-Filho JS, Kumar-Sinha C, Chinnaiyan AM. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. Nat Med. 2011 Nov 20;17(12):1646-51.

3. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. Transcriptome sequencing to detect gene fusions in cancer. Nature. 2009 Mar 5;458(7234):97-101.

4. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, Lonigro RJ, Schroth G, Kumar-Sinha C, Chinnaiyan AM. Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A. 2009 Jul 28;106(30):12353-8.

5. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009 May 1;25(9):1105-11.

6. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010 Sep 1;26(17):2204-7.

7. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.

8. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5.

9. Ghosh D. Discrete nonparametric algorithms for outlier detection with genomic data. J Biopharm Stat. 2010 Mar;20(2):193-208.

Appendices

BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 20 2011, pages 2903–2904
doi:10.1093/bioinformatics/btr467

Genome analysis

Advance Access publication August 11, 2011

ChimeraScan: a tool for identifying chimeric transcription in sequencing data

Matthew K. Iyer[1,2], Arul M. Chinnaiyan[1,2,3,4,5] and Christopher A. Maher[1,2,3,*]

[1]Michigan Center for Translational Pathology, [2]Center for Computational Medicine and Biology, [3]Department of Pathology, [4]Howard Hughes Medical Institute and [5]Department of Urology, University of Michigan Medical School, Ann Arbor, MI 48109, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Next generation sequencing (NGS) technologies have enabled *de novo* gene fusion discovery that could reveal candidates with therapeutic significance in cancer. Here we present an open-source software package, ChimeraScan, for the discovery of chimeric transcription between two independent transcripts in high-throughput transcriptome sequencing data.

Availability: http://chimerascan.googlecode.com

Contact: cmaher@dom.wustl.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 4, 2011; revised on July 26, 2011; accepted on August 3, 2011

1 INTRODUCTION

High-throughput transcriptome sequencing (RNA-Seq) facilitates detection of aberrant, chimeric RNAs (Maher *et al.*, 2009a; Maher *et al.*, 2009b). Methods for chimera detection have already uncovered recurrent classes of clinically relevant gene fusions in prostate (Palanisamy *et al.*, 2010) and lymphoid cancers (Steidl *et al.* 2011). Therefore, the continued development of accurate and efficient software tools for chimera discovery is of major clinical significance. To this end, we have developed a chimera discovery methodology, or ChimeraScan, and offer it as open-source software package for the community to utilize for their own sequencing efforts. ChimeraScan includes features such as the ability to process long (>75 bp) paired-end reads, processing of ambiguously mapping reads, detection of reads spanning a fusion junction, integration with the popular Bowtie aligner (Langmead *et al.*, 2009), supports the standardized SAM format and generation of HTML reports for easy investigation of results. Overall, we believe that the ChimeraScan will facilitate the discovery of additional gene fusions that may serve as clinically relevant targets in cancer.

2 METHODS

Initial paired-end alignment: ChimeraScan uses Bowtie to align paired-end reads to a combined genome-transcriptome reference. An indexing program creates the combined index from genomic sequences (FASTA format) and transcript features (UCSC GenePred format). Paired alignments within the fragment size range (default: 0–1000) are referred to as concordantly mapping reads (Fig. 1A). ChimeraScan uses these alignments to estimate the insert size distribution of the library, which will later be used to filter out likely false positive chimeras.

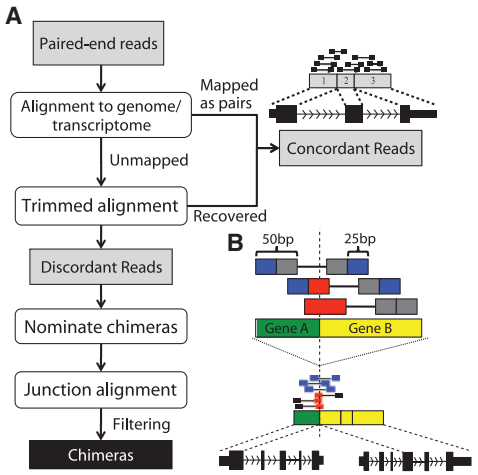

Fig. 1. ChimeraScan flowchart. (**A**) Paired-end reads failing an initial alignment step are segmented and realigned to detect discordant reads. Discordant reads that pass filter criteria are realigned across putative chimeric junctions. (**B**) Chimera with encompassing (blue) and spanning (red) segments detected during realignment.

Trimmed paired-end alignment: read pairs that could not be aligned concordantly are trimmed into smaller segments (default = 25 bp) and realigned. Trimming increases the chance that neither read alignment spans a chimeric junction, thereby improving sensitivity for nominating chimeras.

Nomination of chimera candidates: the trimmed alignments are scanned for evidence of discordant read pairs, or reads that align to distinct references or distant genomic locations (as determined by the fragment size range) of the same reference. Reads aligning to overlapping transcripts are not considered discordant. ChimeraScan clusters the discordant reads and produces a list of putative 5′–3′ transcript pairs that serve as chimera candidates.

Detection of reads spanning the chimeric junction: ChimeraScan builds a new reference index from the set of putative chimeric junction sequences, and realigns candidate junction-spanning reads to this index. Candidate spanning reads are either (i) discordant reads with trimmed alignments bordering a junction or (ii) unmapped reads whose mates align to a predicted chimera (Fig. 1B). A read that spans a junction by more than a minimum 'anchor' length is denoted as a 'spanning' read. We compute the required 'anchor' length separately for each chimera by insisting that the number of bases overlapping its junction be greater than number of homologous bases between the 5′ and 3′ genes at the breakpoint plus the number of mismatches allowed.

*To whom correspondence should be addressed.

Filtering false-positive chimeras: after spanning reads are incorporated, ChimeraScan filters chimeras with few supporting reads (default is <3 reads) and chimeras with fragment sizes far outside the range of the distribution (default is >99% of all fragment sizes). When isoforms of the same gene support a fusion ChimeraScan only retains the isoform(s) with highest coverage.

Reporting chimeras: ChimeraScan produces a tabular text file describing each chimera, and optionally generates a user-friendly HTML page with links to detailed descriptions of the chimeric genes.

3 RESULTS

To evaluate the results from ChimeraScan, we applied it to three well-characterized cancer cell lines known to harbor multiple chimeric transcripts: VCaP (prostate cancer, 2×53 bp) (Tomlins *et al.*, 2005), LNCaP (prostate cancer, 2×34 bp) and MCF7 (breast cancer, 2×35 bp) (Hampton *et al.*, 2009; Volik *et al.*, 2006). Sequence data are deposited in GenBank under the accession number GSE29098. We aligned to human genome (VR-hg19) and UCSC known transcripts (December 2010), allowing for up to two mismatches and no >100 alignments per read. The trimmed alignment step was performed with 25 bp segments.

As our initial benchmark, we confirmed that ChimeraScan was able to recapitulate experimentally validated candidates, our 'gold standard' (Supplementary Table 1) (Maher *et al.*, 2009b). ChimeraScan was able to detect 9/10, 4/4 and 12/13 chimeras from VCaP, LNCaP and MCF-7, respectively.

In addition to recapitulating previously reported results, we have identified novel candidates that demonstrate ChimeraScan's ability to identify and prioritize high-quality chimeras. Overall, ChimeraScan nominated 335 novel chimeras (78 in VCaP, 105 in LNCaP and 152 in MCF7) from the three cell lines (Supplementary Table 2–4). Interestingly, we detected an interchromosomal rearrangement *TBL1XR1-RGS17* detected in the MCF-7 cell line. While not originally reported within NGS data (Maher *et al.*, 2009b), *TBL1XR1-RGS17* was previously detected by a paired-end diTag approach and experimentally confirmed (Ruan *et al.*, 2007). Another novel candidate was the intrachromosomal rearrangement, *NDUFAF2-MAST4*, in VCaP that is supported by just two encompassing reads and one spanning reads. The ability to identify a high-quality spanning read that uniquely confirms the fusion junction (Supplementary Table 2), thereby increasing our confidence in *NDUFAF2-MAST4*, demonstrates the sensitivity of ChimeraScan.

We next compared ChimeraScan with publicly available tools deFuse (McPherson *et al.*, 2011), shortFuse (Kinsella *et al.*, 2011) and MapSplice (Wang *et al.*, 2010) using the 10 experimentally validated VCaP chimeras (Supplementary Table 5). While deFuse nominated the fewest chimeras, it only detected 60% of the true positives. In comparison, ChimeraScan detected 90% of the true positives from 78 predicted chimeras. Of the remaining programs, MapSplice nominated 400 chimeras while detecting 60% of the true positives and ShortFuse nominated 245 chimeras while confirming 70% of the true positives. Overall, these results suggest that ChimeraScan is among the more stringent programs while enriching for true positives.

4 CONCLUSION

Here, we present an optimized publicly available chimera discovery methodology for identifying novel therapeutically targetable gene fusions in human cancers. Our results suggest that ChimeraScan produces a stringent list of predictions that are enriched with true positives. Furthermore, due to its trimmed alignment steps we believe ChimeraScan will be scalable when longer reads are available to provide increased coverage of fusion junctions. Overall, we feel that with the existing features ChimeraScan is a user-friendly tool that will enable other research groups to make discoveries within their own RNA-Seq data collections.

Funding: Department of Defense Breast Cancer Predoctoral Grant (to M.K.I.); Prostate Cancer Foundation Young Investigator Award and National Institutes of Health Pathway to Independence (K99 CA149182-01) Award (to C.A.M.); National Institutes of Health, Department of Defense and Early Detection Research Network (to A.M.C.).

Conflict of Interest: none declared.

REFERENCES

Hampton,O.A. *et al.* (2009) A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res.*, **19**, 167–177.

Kinsella,M. *et al.* (2011) Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics*, **27**, 1068–1075.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Maher,C.A. *et al.* (2009a) Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.

Maher,C.A. *et al.* (2009b) Chimeric transcript discovery by paired-end transcriptome sequencing, *Proc. Natl Acad. Sci. USA*, **106**, 12353–12358.

McPherson,A. *et al.* (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput. Biol.*, **7**, e1001138.

Palanisamy,N. *et al.* (2010) Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nat. Med.*, **16**, 793–798.

Ruan,Y. *et al.* (2007) Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res.*, **17**, 828–838.

Steidl,C. *et al*. (2011) MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature*, **471**, 377–381.

Tomlins,S.A. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.

Volik,S. *et al.* (2006) Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res.*, **16**, 394–404.

Wang,K. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

nature medicine

Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer

Dan R Robinson[1,2,10], Shanker Kalyana-Sundaram[1–3,10], Yi-Mi Wu[1,2], Sunita Shankar[1,2], Xuhong Cao[1,2,4], Bushra Ateeq[1,2], Irfan A Asangani[1,2], Matthew Iyer[1,5], Christopher A Maher[1,2,5], Catherine S Grasso[1,2], Robert J Lonigro[1,2], Michael Quist[1,2], Javed Siddiqui[1,2], Rohit Mehra[1,2], Xiaojun Jing[1,2], Thomas J Giordano[2,6], Michael S Sabel[6,7], Celina G Kleer[2,6], Nallasivam Palanisamy[1,2], Rachael Natrajan[8], Maryou B Lambros[8], Jorge S Reis-Filho[8], Chandan Kumar-Sinha[1,2] & Arul M Chinnaiyan[1,2,4–6,9]

Breast cancer is a heterogeneous disease that has a wide range of molecular aberrations and clinical outcomes. Here we used paired-end transcriptome sequencing to explore the landscape of gene fusions in a panel of breast cancer cell lines and tissues. We observed that individual breast cancers have a variety of expressed gene fusions. We identified two classes of recurrent gene rearrangements involving genes encoding microtubule-associated serine-threonine kinase (MAST) and members of the Notch family. Both MAST and Notch-family gene fusions have substantial phenotypic effects in breast epithelial cells. Breast cancer cell lines harboring Notch gene rearrangements are uniquely sensitive to inhibition of Notch signaling, and overexpression of *MAST1* or *MAST2* gene fusions has a proliferative effect both *in vitro* and *in vivo*. These findings show that recurrent gene rearrangements have key roles in subsets of carcinomas and suggest that transcriptome sequencing could identify individuals with rare, targetable gene fusions.

Recurrent gene fusions and translocations have long been associated with hematologic malignancies and rare soft-tissue tumors as being 'driving' genetic lesions[1–3]. Over the last few years, it has become apparent that these genetic rearrangements are also present in common solid tumors, including a large subset of prostate cancers[4,5] and smaller subsets of lung cancer, among other types of tumors[6]. Secretory breast cancer, a rare subtype of breast cancer, is characterized by recurrent gene fusions of *ETV6* and *NTRK3* (ref. 7). Although multiple breast cancer genomes have been sequenced[8,9], and complex somatic rearrangements have been observed[10], the driving recurrent gene fusions have not been identified.

We used paired-end transcriptome sequencing on a panel of 89 breast cancer cell lines and tumors (**Supplementary Fig. 1**) and then applied our previously developed chimera discovery pipeline[11,12]. This panel represented a spectrum of breast carcinoma and included 42 estrogen receptor (ER)-positive, 21 v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (ERBB2)-positive and 27 triple negative (ER$^-$, progesterone receptor–negative (PR$^-$) and ERBB2$^-$) samples (**Supplementary Table 1**). Investigation of fusion transcripts led to the identification of 384 expressed gene fusions at an average of nearly five fusions per breast cancer sample, with a slightly higher number of gene fusions in the cell lines compared to the primary tumors (**Supplementary Fig. 1b** and **Supplementary Table 2**). Notably, we found that only *SEC16A-NOTCH1* was recurrent in our compendium, even though several fusion genes appeared in combination with different fusion partners. Overall, we found 24 genes to be recurrent fusion partners (**Supplementary Table 2**). To focus on potentially tumorigenic driver fusions, we prioritized the gene fusions based on the known cancer-associated functions of component genes. Although there were many singleton fusions in our compendium that met these criteria, we identified five instances of fusions of MAST family kinases and eight instances of fusions of genes in the Notch family (**Fig. 1** and **Supplementary Fig. 2**).

The genes encoding members of the MAST kinase family are characterized by the presence of a serine-threonine kinase domain, a second 3′ MAST domain with some similarity to kinase domains and a PDZ domain[13]. Little is known about the biological role of MAST kinases, and somatic alterations have not previously been described in cancer. Initially, we identified three independent instances of MAST gene fusions using transcriptome analyses: fusions of *ARID1A* and *MAST2*, *ZNF700* and *MAST1*, and *NFIX* and *MAST1* (**Fig. 1a**). We devised a targeted sequencing approach to screen additional samples for MAST gene fusions. We generated and captured a transcriptome library of 74 pooled breast carcinoma RNAs with baits encompassing

[1]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, Michigan, USA. [2]Department of Pathology, University of Michigan, Ann Arbor, Michigan, USA. [3]Department of Environmental Biotechnology, Bharathidasan University, Tiruchirappalli, India. [4]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan, USA. [5]Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, Michigan, USA. [6]Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan, USA. [7]Department of Surgery, University of Michigan, Ann Arbor, Michigan, USA. [8]The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, London, UK. [9]Department of Urology, University of Michigan, Ann Arbor, Michigan, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to A.M.C. (arul@umich.edu) or C.K.-S. (chakumar@med.umich.edu).

Received 19 April; accepted 24 October; published online 20 November 2011; doi:10.1038/nm.2580

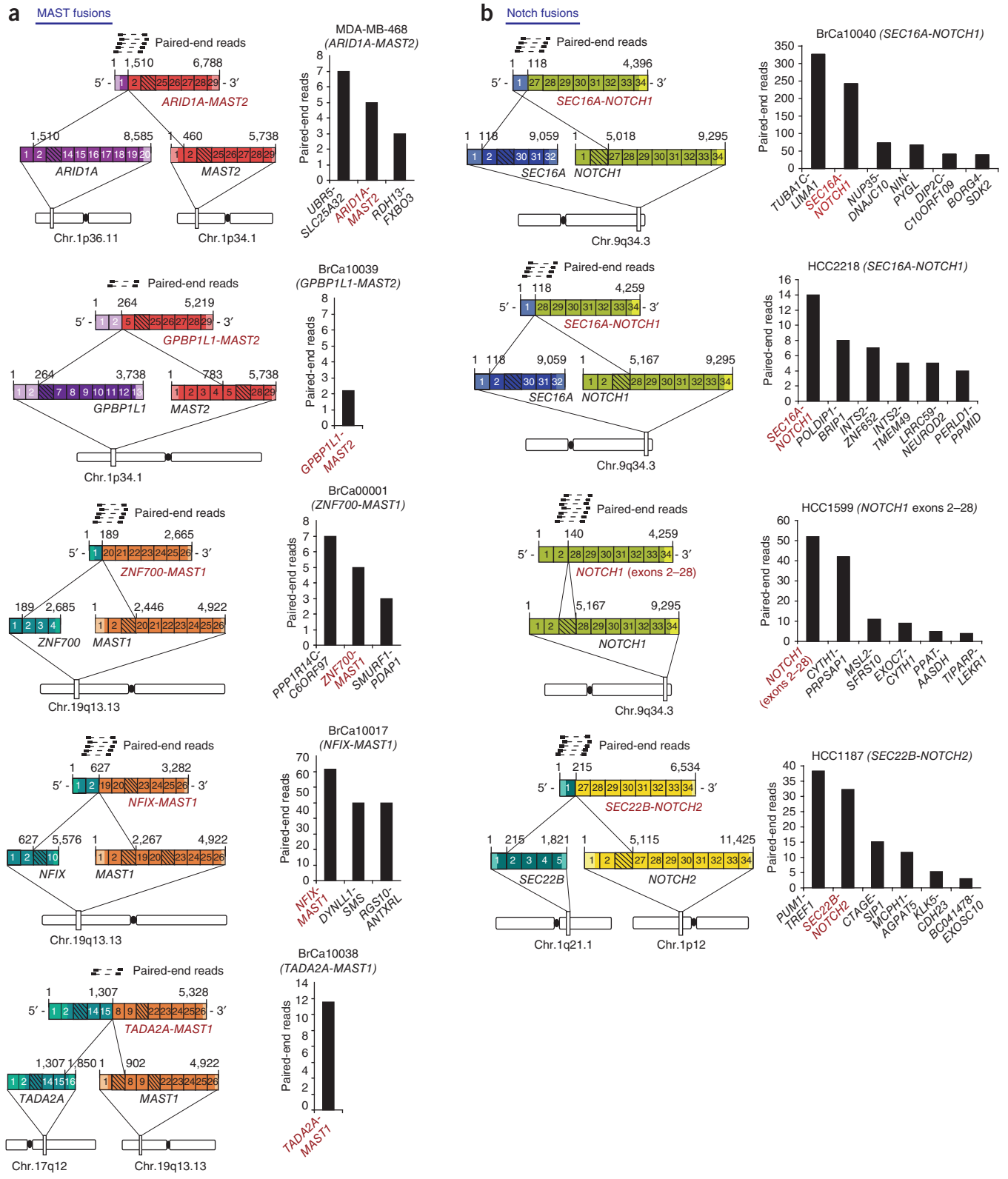

Figure 1 Discovery of the MAST kinase and Notch gene fusions in breast cancer identified by paired-end transcriptome sequencing. (**a**) MAST family gene fusions. (**b**) Notch-family gene fusions. Fusion junctions with respective exon numbers (and nt positions) comprising the chimeric transcripts are shown. Bar plots of the top ranked gene fusions by number of paired-end reads supporting each nominated fusion in the index samples (shown on the right), with MAST or Notch fusion genes shown in red.

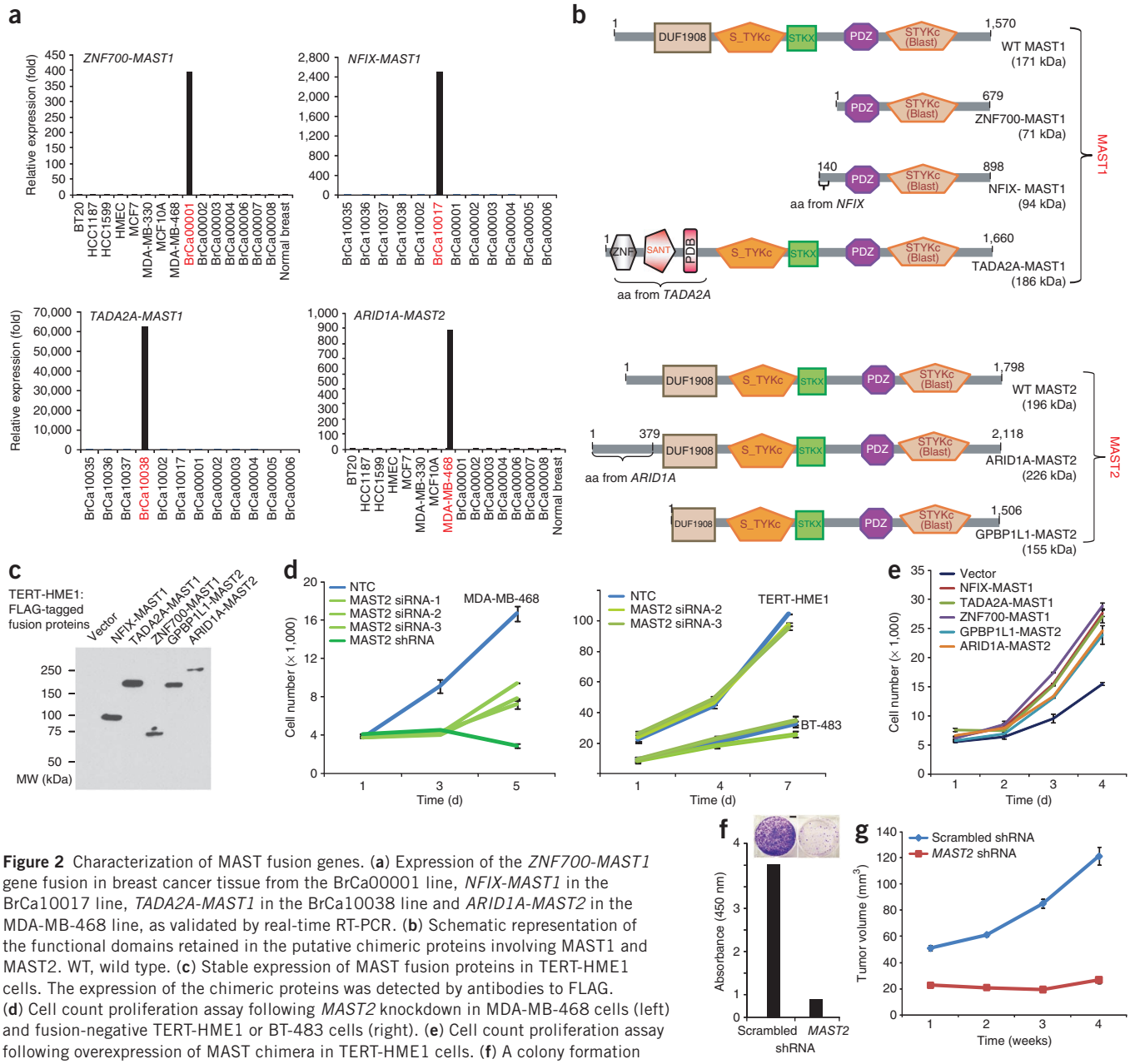

Figure 2 Characterization of MAST fusion genes. (**a**) Expression of the *ZNF700-MAST1* gene fusion in breast cancer tissue from the BrCa00001 line, *NFIX-MAST1* in the BrCa10017 line, *TADA2A-MAST1* in the BrCa10038 line and *ARID1A-MAST2* in the MDA-MB-468 line, as validated by real-time RT-PCR. (**b**) Schematic representation of the functional domains retained in the putative chimeric proteins involving MAST1 and MAST2. WT, wild type. (**c**) Stable expression of MAST fusion proteins in TERT-HME1 cells. The expression of the chimeric proteins was detected by antibodies to FLAG. (**d**) Cell count proliferation assay following *MAST2* knockdown in MDA-MB-468 cells (left) and fusion-negative TERT-HME1 or BT-483 cells (right). (**e**) Cell count proliferation assay following overexpression of MAST chimera in TERT-HME1 cells. (**f**) A colony formation assay of MDA-MB-468 cells treated with *MAST2*-specific shRNA or control scrambled shRNA. The inset shows crystal violet staining of cells treated with either scrambled or *MAST2* shRNA. Scale bar, 5 mm. (**g**) Tumor volumes of MDA-MD-468 xenografts with *MAST2* knockdown. MW, molecular weight; aa, amino acids. Error bars represent s.d.

MAST1 and *MAST2*. After sequencing, we discovered two new MAST gene fusions: *TADA2A-MAST1* and *GPBP1L1-MAST2* (**Fig. 1a**). The samples with MAST gene fusions are distinct from those with Notch family gene fusions (**Fig. 1b**).

We investigated the function of the MAST fusions (**Fig. 2**) and confirmed the fusions using fusion-specific PCR (**Fig. 2a**). All five MAST fusions encoded contiguous open reading frames (ORFs), some of which retained the canonical serine-threonine kinase domain and all of which retained the PDZ domain and the 3′ kinase-like domain (**Fig. 2b**). Therefore, in total, we discovered five new gene fusions encoding MAST1 and MAST2 in a cohort of approximately 100 breast cancer samples and more than 40 cell lines, suggesting that the newly identified MAST gene fusions are present in a subset of 3–5% of breast cancers.

The *ZNF700-MAST1* fusion transcript encodes a truncated MAST1 protein that retains the 3′ kinase-like and PDZ domains. We cloned the ORF of the *ZNF700-MAST1* fusion gene to test its phenotypic effects and used a full-length *MAST2* expression construct to mimic the function of *ARID1A-MAST2* overexpression. To assess the potential oncogenic functions of genes encoding MAST, we ectopically overexpressed epitope-tagged truncated MAST1 and full-length MAST2 in the benign breast cell line TERT-HME1 (**Supplementary Fig. 3a–h**). We then cloned and expressed all five *MAST1* and *MAST2* fusions. Consistent with the earlier observations, TERT-HME1 cells overexpressing the five MAST fusions (**Fig. 2c**) had greater cell proliferation (**Fig. 2e**). Overall, these results suggest that ectopic expression of the MAST fusions results in growth and a proliferative advantage in benign breast epithelial cells.

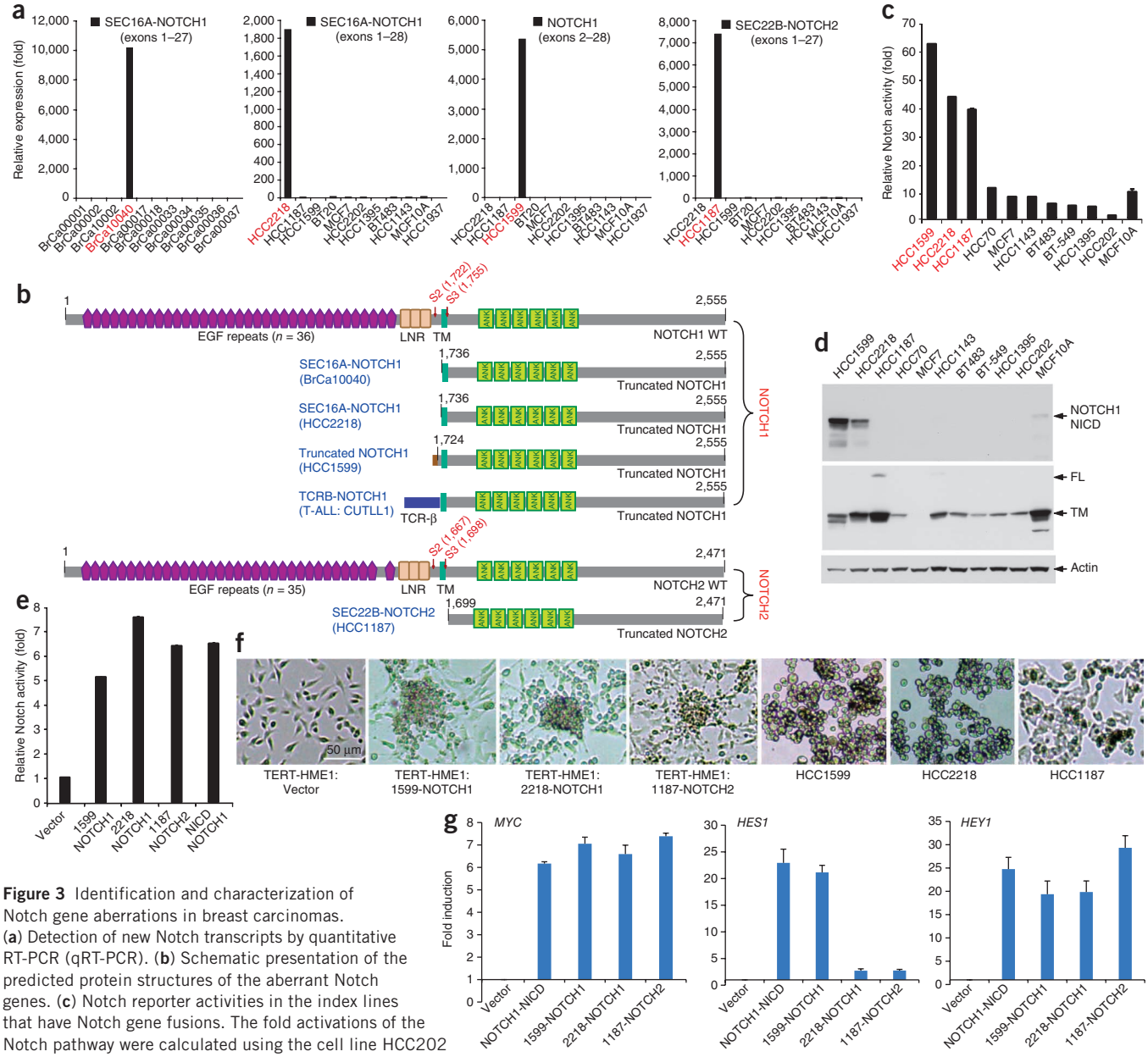

Figure 3 Identification and characterization of Notch gene aberrations in breast carcinomas. (**a**) Detection of new Notch transcripts by quantitative RT-PCR (qRT-PCR). (**b**) Schematic presentation of the predicted protein structures of the aberrant Notch genes. (**c**) Notch reporter activities in the index lines that have Notch gene fusions. The fold activations of the Notch pathway were calculated using the cell line HCC202 as the reference. All the data was normalized to *Renilla* luciferase activity. (**d**) Western blot analysis of NOTCH1-NICD expression, detected with an antibody specifically recognizing the active NOTCH1-NICD protein after cleavage by γ-secretase. FL, full length NOTCH1; TM, transmembrane NOTCH1. (**e**) Notch reporter luciferase levels in 293T cells, assayed following transient Notch expression. (**f**) Bright-field images of control vector and TERT-HME1 cells expressing the Notch fusion alleles. Breast cancer lines positive for Notch fusion are shown for comparison. (**g**) mRNA expression levels of three Notch target genes inTERT-HME1 cells stably expressing Notch fusions, measured by quantitative RT-PCR. Error bars represent s.d.

Because the endogenous *ARID1A-MAST2* fusion is present in the breast cancer cell line MDA-MB-468, we used multiple independent siRNAs specific to *MAST2* or the *ARID1A-MAST2* fusion to achieve knockdown of the ARID1A-MAST2 fusion protein (**Supplementary Fig. 3i–s**). Knockdown of MAST2 showed significant inhibitory effects on growth in MDA-MB-468 cells but not in the fusion-negative cell line BT-483 or in benign TERT-HME1 breast cells (**Fig. 2d**). To further characterize the effects of the *ARID1A-MAST2* fusion in MDA-MB-468 cells, we used shRNA targeting *MAST2*, which showed efficient knockdown of *ARID1A-MAST2* fusion transcript and protein (**Supplementary Fig. 3k,l**). MDA-MB-468 cells treated with *MAST2* shRNA had a reduction in growth, as shown in a colony formation assay (**Fig. 2f**), and showed increased apoptosis and S-phase arrest (**Supplementary Fig. 3m,n**). In the mouse xenograft model, MDA-MB-468 cells transiently transfected with *MAST2* shRNA did not establish palpable tumors over a time course of 4 weeks after transfection (**Fig. 2g**). Our knockdown studies showed that the *ARID1A-MAST2* fusion is a key driver fusion in MDA-MB-468 cells.

In addition to MAST fusions, we found a total of eight rearrangements involving either *NOTCH1* or *NOTCH2* (**Fig. 1b** and **Supplementary Fig. 2**). We found all of these rearrangements in ER$^-$ breast carcinomas ($P = 0.008$) and all but one rearrangement in triple-negative breast carcinomas. We focused on one ER$^-$ tumor and three ER$^-$ breast cancer cell lines with 3′ *NOTCH1* or *NOTCH2*

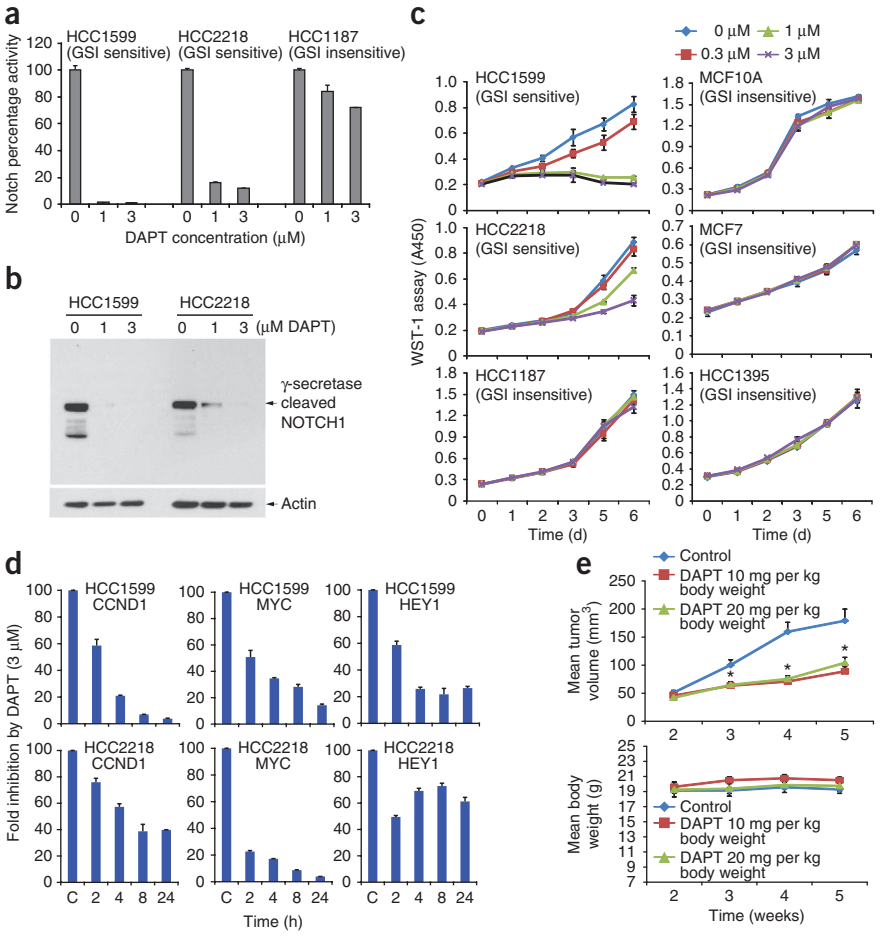

Figure 4 γ-secretase inhibitor DAPT effects in fusion positive and negative breast carcinoma cell lines. (**a**) Luciferase assay of the Notch signaling pathway following DAPT treatment. Breast cancer cells were co-infected with a Notch reporter construct, lenti-RBPJ (recombination signal binding protein for immunoglobulin κJ) firefly luciferase, and the internal control lenti-*Renilla* luciferase. Twenty-four hours after treatment with DAPT, luciferase activities were measured. (**b**) NICD levels after treatment with DAPT detected using an antibody specific to active NOTCH1-NICD after cleavage by γ-secretase. (**c**) WST-1 cell proliferation assays of six breast cell lines after DAPT treatment. (**d**) Expression of Notch target genes after treatment with DAPT, as measured by qRT-PCR. (**e**) Xenograft tumor volume and body mass after treatment with the γ-secretase inhibitor DAPT. Mice xenografted with HCC1599 cells were treated daily after tumors formed, and the size of the tumors was monitored. $^*P < 0.005$.

fusion transcripts in our functional studies. The Notch fusion transcripts were abundantly expressed and were specific to the samples with DNA rearrangements (**Fig. 3a**). All the fusion transcripts retained the exons that encode the Notch intracellular domain (NICD), which is responsible for inducing the transcriptional program following Notch activation (**Fig. 3b**). We characterized the DNA breakpoints associated with Notch fusions by mate-pair genomic library sequencing or by long-range genomic PCR (**Supplementary Fig. 4a**,**b**).

We categorized the predicted ORFs for the *NOTCH1* and *NOTCH2* fusion transcripts into two classes (**Fig. 3b**). For both the *SEC16A-NOTCH1* fusions and the intragenic *NOTCH1* fusion in the HCC1599 cell line, the predicted ORFs initiated after the S2 cleavage site but before the S3 γ-secretase cleavage site, similar to that seen in the *TCRB-NOTCH1* fusion in the adult lymphocytic leukemia T cell line CUTLL1 (ref. 14). In contrast, we predicted the *SEC22B-NOTCH2* fusion ORF to initiate just after the γ-secretase S3 cleavage site. The resulting protein would be nearly identical to NICD, and we predict that it would be highly active and independent of cleavage by γ-secretase (**Fig. 3b**).

We saw substantially higher Notch responsive transcriptional activity in the three cell lines with Notch fusions compared to the other breast cell lines using a Notch luciferase reporter (**Fig. 3c**). Therefore, each of the three Notch fusions is capable of activating the expression of Notch-responsive genes. Using an antibody specific to the γ-secretase cleaved active form of the NOTCH1 NICD, both HCC1599 and HCC2218 showed high concentrations of NICD, consistent with the fusion protein acting as a substrate for activation by γ-secretase (**Fig. 3d**). The HCC1187 cell line, which has a *NOTCH2* fusion gene, contains little NOTCH1 NICD. Most breast cancer lines express wild-type *NOTCH1* (**Fig. 3d**, middle); however, only the two cell lines with *NOTCH1* fusion alleles showed high concentrations of activated NICD. Each of the three fusion alleles, which we co-transfected with a Notch reporter plasmid, induced Notch-responsive transcription that was equivalent to NICD (**Fig. 3e**).

The three breast cell lines containing the Notch fusions showed decreased cell-matrix adhesion and grew in suspension or as weakly adherent clusters, which was in contrast to the majority of breast carcinoma cell lines. When we transduced *NOTCH1* and *NOTCH2* fusion alleles to create stable pools of TERT-HME1 cells, we observed notable morphological changes (**Fig. 3f**). TERT-HME1 cells had adherent epithelial properties, whereas cells expressing Notch fusion lost adherence and propagated as weakly attached clusters, similar to the index lines with Notch fusions and consistent with the previously reported effects of NICD expression in MCF10A cells[15]. Furthermore, the fusion alleles markedly induced expression of the Notch target genes *MYC*, *HES1* and *HEY1* (**Fig. 3g**).

The Notch fusions represent two functional classes with respect to dependence on the activity of γ-secretase. Fusions in BrCa10040, HCC2218 and HCC1599 cells are dependent on S3 cleavage for activity and are sensitive to γ-secretase inhibitors (GSIs). The fusion class in HCC1187 cells is independent of S3 cleavage. We established stable Notch reporter lines from each of the three Notch fusion index lines and treated them with the γ-secretase inhibitor *N*-[(3,5-difluorophenyl)acetyl]-L-al anyl-2-phenyl]glycine-1,1-dimethylethyl ester (DAPT)[16]. We saw a reduction of Notch reporter activity after treatment with DAPT in the HCC1599 and HCC2218 fusion alleles (**Fig. 4a**). However, Notch reporter activity was only slightly diminished by treatment with DAPT in HCC1187 cells, which express a γ-secretase–independent Notch fusion allele that is capable of activating Notch reporter activity. DAPT treatment also substantially reduced NICD protein concentrations in both of the γ-secretase inhibitor–sensitive cell lines (**Fig. 4b**). Furthermore, the index cell lines showed dependence on Notch signaling for proliferation and survival

(**Fig. 4c**). The HCC1599 and HCC2218 cell lines showed marked reductions in proliferation after treatment with DAPT. The HCC1187 cell line, which expresses GSI-independent *NOTCH2* fusion, had no reduction in proliferation after DAPT treatment, which is also the case in breast cell lines not expressing Notch fusion alleles.

Treatment with DAPT repressed the expression of the Notch targets *MYC* and *CCND1* (**Fig. 4d**), two genes that have a key role in mouse mammary tumorigenesis induced by Notch[17,18], which further supports the idea GSIs could be useful in treating cancers that have activated Notch alleles. Consistent with this, treatment with DAPT significantly reduced tumor volume in a xenograft tumor model of HCC1599 cells (**Fig. 4e**).

Since the discovery of the *TMPRSS2-ERG* gene fusion in approximately 50% of prostate cancers, emerging evidence has suggested that recurrent gene fusions have a more substantial role in common solid tumors than was previously known. The MAST and Notch aberrations in breast cancer are new classes of rare but functionally recurrent gene fusions with therapeutic implications (similar to the anaplastic lymphoma receptor tyrosine kinase (ALK) fusions in lung cancer). MAST kinase and Notch gene rearrangements were mutually exclusive aberrations in the samples we tested, and, together, may be present in up to 5–7% of breast cancers. The discovery of functionally recurrent MAST and Notch fusions in a subset of breast carcinomas is a promising path for future research and treatment in breast cancer and illustrates the power of next-generation sequencing as a tool in the development of personalized medicine.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemedicine/.

Note: Supplementary information is available on the Nature Medicine website.

ACKNOWLEDGMENTS

We thank R. Morey for high-throughput sequencing support; T. Barrette for hardware and database management; R. Wang, N. Consul, C. Malla, L. Ma, J. Milton, L. Cai and M. Mei for technical help; and K. Suleman and W. Yan for help with cytogenetic analysis. D. Appledorn from Essen Bioscience performed the IncuCyte analyses. The aims of this project were defined by the Department of Defense Breast Cancer Research Program (W81XWH-08-0110) to A.M.C. This project was supported in part by an American Association for Cancer Research Stand Up to Cancer (SU2C) award to A.M.C. and J.S.R.-F., grant R01 CA125577 to C.G.K. The National Functional Genomics Center (W81XWH-11-1-0520), which is supported by the Department of Defense (A.M.C.) and, in part, by the US National Institutes of Health through the University of Michigan's Cancer Center Support grant 5 P30 CA46592. A.M.C. is supported by the US National Cancer Institute's Early Detection Research Network (U01 CA111275), the Doris Duke Charitable Foundation Clinical Scientist Award and the Burroughs Welcome Foundation Award in Clinical Translational Research. R.N., M.B.L. and J.S.R.-F. are funded in part by Breakthrough Breast Cancer. A.M.C. is an American Cancer Society Research professor and Taubman Scholar. C.K.S. and A.M.C. share senior authorship.

AUTHOR CONTRIBUTIONS

D.R.R., C.K.-S. and A.M.C. conceived of the experiments. D.R.R., C.K.-S., Y.-M.W. and X.C. performed transcriptome sequencing. D.R.R., Y.-M.W. and X.C. performed target capture screening and sequencing. S.K.-S., C.A.M. and M.I. performed the bioinformatics analysis of high-throughput sequencing data and the nomination of gene fusions. C.S.G., R.J.L. and M.Q. performed bioinformatic analysis of high-throughput sequencing data for the gene expression profiling. C.K.-S., D.R.R. and Y.-M.W. performed the gene fusion validations. S.S. performed the *in vitro* experiments of MAST. I.A.A. performed the chorioallantoic membrane assays. B.A. performed the xenograft experiments. D.R.R. and Y.-M.W. performed the *in vitro* experiments of Notch. X.J. performed the microarray experiments. J.S., M.S.S., C.G.K., T.J.G., N.P., R.N., M.B.L. and J.S.R.-F. provided breast cancer tissue samples and the associated clinical annotation. N.P. performed fluorescence *in situ* hybridization experiments, and R.M. evaluated the fluorescence *in situ* hybridization results. D.R.R., C.K.-S. and A.M.C. wrote the manuscript, which was reviewed by all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemedicine/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Delattre, O. *et al.* Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. *Nature* **359**, 162–165 (1992).
2. Nowell, P.C. & Hungerford, D.A. Chromosome studies on normal and leukemic human leukocytes. *J. Natl. Cancer Inst.* **25**, 85–109 (1960).
3. Rowley, J.D. The critical role of chromosome translocations in human leukemias. *Annu. Rev. Genet.* **32**, 495–519 (1998).
4. Kumar-Sinha, C., Tomlins, S.A. & Chinnaiyan, A.M. Recurrent gene fusions in prostate cancer. *Nat. Rev. Cancer* **8**, 497–511 (2008).
5. Tomlins, S.A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
6. Prensner, J.R. & Chinnaiyan, A.M. Oncogenic gene fusions in epithelial carcinomas. *Curr. Opin. Genet. Dev.* **19**, 82–91 (2009).
7. Tognon, C. *et al.* Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. *Cancer Cell* **2**, 367–376 (2002).
8. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
9. Shah, S.P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
10. Stephens, P.J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).
11. Maher, C.A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97–101 (2009).
12. Maher, C.A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 12353–12358 (2009).
13. Garland, P., Quraishe, S., French, P. & O'Connor, V. Expression of the MAST family of serine/threonine kinases. *Brain Res.* **1195**, 12–19 (2008).
14. Palomero, T. *et al.* CUTLL1, a novel human T-cell lymphoma cell line with t(7;9) rearrangement, aberrant NOTCH1 activation and high sensitivity to gamma-secretase inhibitors. *Leukemia* **20**, 1279–1287 (2006).
15. Mazzone, M. *et al.* Dose-dependent induction of distinct phenotypic responses to Notch pathway activation in mammary epithelial cells. *Proc. Natl. Acad. Sci. USA* **107**, 5012–5017 (2010).
16. Dovey, H.F. *et al.* Functional gamma-secretase inhibitors reduce beta-amyloid peptide levels in brain. *J. Neurochem.* **76**, 173–181 (2001).
17. Klinakis, A. *et al.* Myc is a Notch1 transcriptional target and a requisite for Notch1-induced mammary tumorigenesis in mice. *Proc. Natl. Acad. Sci. USA* **103**, 9262–9267 (2006).
18. Ling, H., Sylvestre, J.R. & Jolicoeur, P. Notch1-induced mammary tumor development is cyclin D1-dependent and correlates with expansion of pre-malignant multipotent duct-limited progenitors. *Oncogene* **29**, 4543–4554 (2010).

ONLINE METHODS

Cell lines and specimen collection. Breast cancer cell lines were purchased from the American Type Culture Collection. The tissue was collected under approval of the University of Michigan Institutional Review Board IRBMED under approved protocol HUM00041989, and breast cancer samples were obtained with informed consent at the University of Michigan and the Breakthrough Breast Cancer Research Centre, Institute of Cancer Research (London, UK).

Paired-end transcriptome sequencing. Total RNA was extracted from healthy and cancer breast cell lines and breast tumor tissues, and the quality of the RNA was assessed with the Agilent Bioanalyzer. Transcriptome libraries from the mRNA fractions were generated following the RNA-Seq protocol (Illumina). Each sample was sequenced in a single lane with the Illumina Genome Analyzer II (with a 40- to 80-nt read length) or with the Illumina HiSeq 2000 (with a 100-nt read length). Paired-end transcriptome reads passing our filters were mapped to the human reference genome (hg18) and to UCSC genes using Illumina Efficient Alignment of Nucleotide Databases (ELAND) software. Sequence alignments were then processed to nominate gene fusions using a previously described method[11,12].

qRT-PCR and long-range PCR. qRT-PCR assays using SYBR Green Master Mix (Applied Biosystems) were carried out with the StepOne Real-Time PCR System (Applied Biosystems). Relative mRNA levels of each chimera were normalized to the expression of *GAPDH*. To detect the genomic fusion junction in HCC1187 cells, primers were designed that flanked the predicted fusion position, and PCR reactions were performed to amplify the fusion fragments. Oligonucleotide primer sequences are listed in **Supplementary Table 3**.

Immunoblot detection of the MAST2 fusion protein and NOTCH1. An immunoblot analysis of MAST2 was performed using an antibody to MAST2 obtained from Novus Biologicals. Antibody to human β-actin (Sigma-Aldrich) was used as a loading control. For the detection of NOTCH1, cells were lysed in radioimmunoprecipitation assay buffer containing protease inhibitor cocktail (Pierce). Proteins were separated by SDS-PAGE, transferred to nitrocellulose membranes and probed with antibodies recognizing total NOTCH1 (Cell Signaling), γ-secretase–cleaved NOTCH1 (NICD; Cell Signaling) or β-actin (Santa Cruz).

Constructs used for overexpression studies. The *ZNF700-MAST1* fusion ORFs from the BrCa00001 cell line were cloned into a Gateway pcDNA-DEST40 mammalian expression vector (Invitrogen) using LR Clonase II. A plasmid with a C-terminus V5 tag was generated and tested for protein expression after transfection into HEK293 cells. A full-length expression construct of *MAST2* with a DDK tag was obtained from OriGene.

Establishment of stable pools of TERT-HME1 cells. The five MAST fusion alleles were cloned with an N-terminal Flag epitope tag into the lentiviral vector pCDH510-B (SABiosciences). The lentivirus was produced by cotransfecting each of the MAST plasmids using the ViraPower packaging mix (Invitrogen) into 293T cells using FuGENE HD transfection reagent (Roche). Thirty-six hours after transfection, the viral supernatants were collected, centrifuged and then filtered through a 0.45-μm Steriflip filter unit (Millipore). TERT-HME1 cells were infected at a multiplicity of infection of 20 with polybrene at 8 μg ml^{-1}. Forty-eight hours after infection, the cells were split and placed into puromycin-selective medium. Stable pools of TERT-HME1 cells expressing the NOTCH fusion alleles as well as a control NOTCH1 intracellular domain were generated using the same procedures.

Knockdown assay. For siRNA knockdown experiments, multiple independent MAST2 siRNAs from Thermo were used (J-004633-06, J-004633-07 and J-004633-08). All siRNA transfections were performed using Oligofectamine reagent (Life Sciences). Similar experiments were performed with multiple custom siRNA sequences targeting the ARID1A-MAST2 fusion (Thermo). Lentiviral particles expressing the MAST2 shRNA (Sigma, TRCN0000001733) were transduced using polybrene according to the manufacturer's instructions.

Colony formation assay. MDA-MB-468 cells transduced with scrambled or *MAST2* shRNA lentivirus particles were plated and selected using puromycin. After 7–8 d, the plates were stained with crystal violet to visualize the number of colonies formed. For quantification of the differential staining, the plates were treated with 10% acetic acid, and absorbance was read at a wavelength of 750 nm.

Mouse xenograft models. Four-week-old female severe compromised immunodeficiency C.B17 mice were procured from a breeding colony at University of Michigan that is maintained by K. Pienta. Mice were anesthetized using a cocktail of xylazine (80 mg per kg of body weight intraperitoneally (i.p.)) and ketamine (10 mg per kg of body weight i.p.) for chemical restraint. Breast cancer cells with *MAST2* shRNA or scrambled shRNA knockdown ($n = 4$ million) or the HCC1599 breast cancer cell line positive for the *NOTCH1* fusion allele ($n = 5$ million) were resuspended in 100 μl of 1× PBS with 20% Matrigel (BD Biosciences) and implanted into the right and left abdominal inguinal mammary fat pads of the mice. Ten mice were included in each group. Two weeks after tumor implantation, HCC1599 xenografted mice were treated daily with the γ-secretase inhibitor DAPT, which was dissolved in 5% ethanol and corn oil (i.p.). All procedures involving mice were approved by the University Committee on Use and Care of Animals of the University of Michigan.

Additional methods. Detailed methodology is described in the **Supplementary Methods**.

doi:10.1038/nm.2580